EVALUATION OF PARETO/D/1/K QUEUE BY SIMULATION

Seferin Mirtchev, Rossitza Goleva

Abstract: The finding that Pareto distributions are adequate to model Internet packet interarrival times has motivated the proposal of methods to evaluate steady-state performance measures of Pareto/D/1/k queues. Some limited analytical derivation for queue models has been proposed in the literature, but their solutions are often of a great mathematical challenge. To overcome such limitations, simulation tools that can deal with general queueing system must be developed. Despite certain limitations, simulation algorithms provide a mechanism to obtain insight and good numerical approximation to parameters of queues. In this work, we give an overview of some of these methods and compare them with our simulation approach, which are suited to solve queues with Generalized-Pareto interarrival time distributions. The paper discusses the properties and use of the Pareto distribution. We propose a real time trace simulation model for estimating the steady-state probability showing the tail-raising effect, loss probability, delay of the Pareto/D/1/k queue and make a comparison with M/D/1/k. The background on Internet traffic will help to do the evaluation correctly. This model can be used to study the long-tailed queueing systems. We close the paper with some general comments and offer thoughts about future work.

Keywords: Pareto distribution, delay system, queueing analyses, simulation model, peak traffic modelling;

ACM Classification Keywords: G.3 Probability and statistics: queueing theory, I.6.5 Model development

Conference: The paper is selected from Sixth International Conference on Information Research and Applications – i.Tech 2008, Varna, Bulgaria, June-July 2008

Introduction

Managed IP networks have become a dominant factor in bringing information to users on a worldwide basis. Until recently, IP networks supported only a best effort service. This limitation has not been a problem for traditional Internet applications like web and email, but it does not satisfy the needs of many new applications like audio and video streaming, which demand high data throughput capacity (bandwidth) and have low-latency requirements. Thus, it is becoming increasingly important to provide Quality of Service (QoS) in managed IP networks.

As pointed out by several authors who have been collecting traffic data from the Internet, there is no a queueing theory method for queue analyses when one is given a set of packet interarrival times. Obviously, one could fit the resulting data to a distribution and then use a queueing model if it exists. There are some papers concerning batch arrivals like [Khadjiivanov, 1993]. Traffic growth and its influence to the congestion management is demonstrated in [Tsankov, 2007]. Internet traffic can be described as having one or more of the following related characteristics [Cao, 2004, Salvador, 2004]: Self-similar (or fractal) traffic traces; Long-range dependence; Burstiness on multiple scales; Long- or heavy-tailed packet interarrival times or service requirements.

There has been a substantial amount of literature on analyzing and characterizing the traffic appearing on the Internet. The Internet traffic data are well known to possess extreme variability and bursty structure in a wide range of time scales. This characteristic is not found on the Poisson process. The properties can be characterized by self-similar process. The large variation pertaining to the self-similar nature of data traffic causes congestion problems in the data network. The arrival process with Pareto distributed interarrival time is a popular model of self-similar processes.

The queue performance of Pareto/M/1/k was studied by simulations in [Koh, 2003]. They are investigated the queue behaviour with Pareto interarrival distribution. By numerical analysis and simulations, they have been analyzed the asymptotic and the exact loss probabilities of GI/M/1/k to show the big discrepancy between the asymptotic and the actual loss probability and propose a model for the loss probability of Pareto/M/1/k as a function of the buffer size and the geometric parameter.

The Pareto distribution is a model for nonnegative data with a power law probability tail. A natural upper bound truncates the probability tail in many practical applications. An estimators are derived for the truncated Pareto distribution in [Inmaculada, 2006]. They investigate distribution properties and illustrate its applicability in practice.

The simulation of systems using heavy-tailed distributions presents difficulties and needs efficient methods to study. In [Argibay, 2003] there is a trial to go into insight nature of simulation difficulties of M/G/n queues with G heavy-tailed distribution. They have proposed and developed a method to speed up simulations and used M/G/1 systems as workbenches since they have some analytical results to check the results.

Stochastic simulation has become a well established paradigm used in performance evaluation of various complex dynamic systems. In [Eickhoff, 2006] a method for estimating time evolution of several quantiles within some time interval is described. It is based on independent replications and its capability is demonstrated by simulating processes with different kinds of stationary, non-stationary or transient behaviour.

The concept of self-similarity (or fractal behaviour) is the best understood by looking at [Fernandes, 2003]. The use of synthetic self-similar traffic in computer networks simulation is of vital importance for the capturing and reproducing of actual Internet data traffic behaviour. Fernandes uses a technique for self-similar traffic generation that is achieved by aggregating On/Off sources where the active (On) and idle (Off) periods exhibit heavy tailed distributions. This work analyzes the balance between accuracy and computational efficiency in generating self-similar traffic and presents important results that can be useful to parameterize existing heavy tailed distributions such as Pareto, Weibull and Lognormal in a simulation analysis.

The Pareto distribution is a special heavy tailed distribution called a power-tailed distribution. It is found to serve as adequate model for many situations. Gross and al. [Gross, 2003] investigated many difficulties in simulating queues with Pareto service. They considered truncated Pareto service.

A method for studying Pareto queues is presented in [Fischer, 1999]. The paper discusses the properties and use of the Pareto distribution. The method is used to study the Pareto/M/1 queue and look at the M/Pareto/1 queue. The first could be used to model arrivals of packets at a packet switched network, and the second, the time to transmit files through such a network.

The Pareto distribution has various forms. A one and two-parameter form is considered in [Fischer, 2005]. The two Pareto forms are studied in detail. It is shown that the usage of the two-parameter Pareto results in lower congestion than the comparable one-parameter Pareto.

Some limited analytical derivation for queueing models whit Pareto distribution is proposed in the literature, but their solutions are often of a great mathematical challenge. To overcome such limitations, simulation tools that can deal with general queueing systems have to be developed. Despite certain limitations, simulation algorithms provide a mechanism to obtain insight and good numerical approximation to parameters of networks of queues.

This paper presents a stochastic simulation method for studying Pareto queues. The paper discusses the properties and use of the Pareto distribution. We make the comparison between Pareto/D/1/K and M/D/1/K and propose a real time trace simulation model for estimating the steady-state probability showing the tail-raising effect, the loss probability and delay. The background on Internet traffic will help to do the evaluation correctly. This model can be used to study the long-tailed queueing systems.

Generalised Pareto distribution

The most common choice for telecommunication network design is based on the exponential assumption. Usual choice is the Poisson arrival of the calls or sessions and exponential holding times. However, networks and applications of today generate a traffic that is bursty over a wide range of time scales. A number of empirical studies have shown that the network traffic is self-similar or fractal in nature.

The Pareto distribution, named after the Italian economist Vilfredo Pareto, is a power law probability distribution that coincides with social, scientific, geophysical, actuarial, and many other types of observable phenomena.

The family of Generalized Pareto Distributions (GPD) has three parameters: the location parameter μ, the scale parameter σ and the shape parameter ξ.

The cumulative distribution function of the GPD is:

$$F(x) = 1 - \left(1 + \frac{\xi(x-\mu)}{\sigma}\right)^{-1/\xi} . \tag{1}$$

We choose these substitutions

$$\eta_0 = \frac{\xi}{\sigma}; \quad \lambda = \frac{\sigma}{\sigma^2 - \xi}; \quad \mu = 0 . \tag{2}$$

Therefore, we receive another form of the generalized-Pareto distribution:

$$F(t) = 1 - (1 + \eta_o t)^{-\left(1 + \lambda/\eta_o\right)} \tag{3}$$

The mean value of the generalized-Pareto distribution is:

$$m_o = 1/\lambda \tag{4}$$

The mean value is the average interarrival time for our study. The parameter λ is the call arrival intensity.

The variance of the Generalized Pareto Distribution is:

$$d_i = \frac{\lambda + \eta_o}{\lambda^2(\lambda - \eta_o)}, \quad 0 \le \eta_o \le \lambda \tag{5}$$

It follows that the probability density function of the GPD is:

$$f(t) = (\eta_o + \lambda)(1 + \eta_o t)^{-\left(2 + \lambda/\eta_0\right)} \tag{6}$$

It is convenient to define the mean value and variance of the arrival stream. We can easy calculate the parameter η_o (the ratio of the shape and scale parameter):

$$\eta_o = \lambda\left(1 - \frac{2}{d_i\lambda^2 + 1}\right) \tag{7}$$

Random number generation

Many programming languages do not yet recognize the Pareto distribution. In the field of telecommunications, the Pareto distribution is widely used to estimate the interarrival and service times.

One can easily generate a random sample from Pareto distribution by using inverse distribution function. Given a random variable U with uniform distribution on the unit interval (0;1), the random variable *x* is Pareto-distributed.

$$x = \frac{U^{-\frac{\eta_o}{\eta_o + \lambda}} - 1}{\eta_o} \tag{8}$$

Uniformly distributed pseudo-random numbers in the space (0,1] are usually referred to as *random numbers*, whereas random numbers following any other distribution are referred to as *random variates* or *stochastic variates.*

Pareto/D/1/k Simulation Model Description

Recall that in standard queueing notation, A/B/C, “A” represents the arrival distribution, “B” the service distribution, and “C” the number of servers. “M” means “memoryless”, which in this context implies Poisson distribution for arrival rates and exponential distribution for service times. Our simulation model is used to study the Pareto/D/1/k queue. It could be used to model arrivals of packets at a packet switched network.

Let us consider a single server queue Pareto/D/1/k with a Pareto input stream, which is defined by arrival intensity λ, variance of the interarival time d_i, constant service time τ and limited waiting room k. This queueing system with peak input stream and constant service time is a non-Markovian model (Figure 1). It is assumed that customers are served in FCFS order.

Simulations are the main tools for studying the performance of telecommunication networks and we will analyse Pareto/D/1/k queue using simulation.

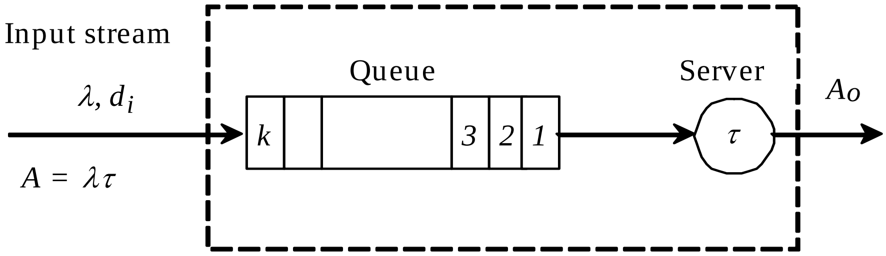

Figure 1. Pareto input stream model with constant service time and finite waiting positions.

The arrival process in Pareto/D/1/k queue is considered to be a renewal process. G/D/1/k queues, where G means a heavy-tailed distribution of interarrival time, are used to model queue systems where a range of values of the interarrival time, whose probability is very low, have a drastic impact on the overall performance of the system. The Pareto distribution is one of these heavy-tailed distributions and it is proposed to describe peak streams in the packet switched networks. The accurate analytical treatment of Pareto/D/1/k systems is very difficult and in many cases it cannot be applied. Simulation is a possible method to study. Simulations with heavy-tailed random variables present some additional difficulties. A care must be taken during analyses of the results of these simulations. It is necessary to have accurate and efficient simulation methods. The efficacy is important because we need to generate big quantities of data for our simulation study and be accurate enough. The data accuracy can be estimated by means of comparisons with known results from simpler systems with analytical solution. One of these simpler queue systems that is studied analytically is the M/D/1/k queue. This queue is used as a workbench for more efficient simulation methods, able to deal with the heavy-tail difficulties.

We develop a real time trace simulation algorithm for evaluating the state probabilities of the queueing system, the call congestion probability and the mean time in the queue. We use batch mean method for output results analysis and choose a confidence probability 95%. We define 20 batches and generate 20000 calls in every batch. We introduce an initial bias to eliminate the influence of the transient behaviour and time intervals between batches to received independent estimates of the call congestion probability and the mean queueing time. We describe the accuracy of the estimates by means of a confidence interval, which with a given probability (95%) specifies how the estimate is placed relatively to the unknown theoretical value, using the Student's t-distribution with 19 degrees of freedom. This organization of our algorithm leads to good accuracy from a practical point of view. The relative errors of the presented results are less than 10%.

Random errors are caused by the stochastic variations of the simulation. They appear because every simulation is similar to a statistical experiment. The next source of error is the bias of the estimator itself, being often called the systematic error. This kind of error usually appears if assumptions about the analyzed data are true only approximately or asymptotically. If both the variance and the bias tend to zero for large number of observations the estimator is called consistent.

Pareto/D/1/k System Performance Measures

In this section, we present some of the parameters used in our analysis and the results of the simulation runs. We will illustrate these parameters by running a different scenario.

OFFERED TRAFFIC

The offered traffic *A* is calculated by means of the average arrival rate and the constant service time

$$A = \lambda \tau \quad . \tag{9}$$

BLOCKING PROBABILITY

The call congestion probability *B* is defined and evaluated by the simulation program as the ratio of lost and arrival calls. It can be calculated by offered and carried traffic

$$B = \left(A - A_o\right) / A \quad . \tag{10}$$

MEAN QUEUEING TIME

The real time simulation gives as possibility to calculate the queueing time for every arrival call and it is easy to obtain the mean queueing time.

Simulation Results

In this section, we give numerical results obtained by a Pascal program on a personal computer. The described models are tested on a computer over a wide range of arguments.

Figure 2 illustrates the stationary probability distribution in a single server queue Pareto/D/1/k (P/D/1/k on the Figures below) with a Pareto input stream, 0.85 erl offered traffic, 30 waiting positions and different variance of the interarrival time. It is seen that when the variance increases the probability that the queue is full increases significantly.

Figure 3 presents the call congestion probability in a single delay system with 30 waiting positions, different offered traffic and different variance of the interarrival time. When the offered traffic is comparatively small (0.8 erl) the influence of the variance of the call congestion probability is great.

Figure 4 shows the queueing time as function of the offered traffic when the number of queueing positions is 30, the service time is 1 second and different variance of the interarrival time.

It is shown that the influence of the variance of the input stream over the performance measures is significant. The heavy-tailed condition decisively contributes to raise the congestion and waiting time.

The computer simulation of Pareto/D/1/k queues presents important difficulties due to the slow decaying tail of the Pareto distribution. This makes extremely high values, with great influence on the statistical figures of the system. The probabilities are so low that in case we want to simulate the physical underlying processes, generating demanded times and time arrivals, the cost in time will probably be prohibitive if we want accurate results. This forces to use all our knowledge of the statistics of the system inner processes, so the simulation can noticeably speed up.

Conclusion

We have presented a simulation method for evaluating the Pareto/D/1/k queueing systems. We have demonstrated its use by presenting numerical results. These results have shown that the Pareto distribution change significantly the queue behaviour. In the case of Pareto/D/1/k, congestion occurred even when the load is sufficiently small. But for that queue, the long-tailed nature of the Pareto helps to clear out congestion when a large interarrival time occurred. Our model can be applied for all Pareto/D/1/k systems independently of the value of the parameters.

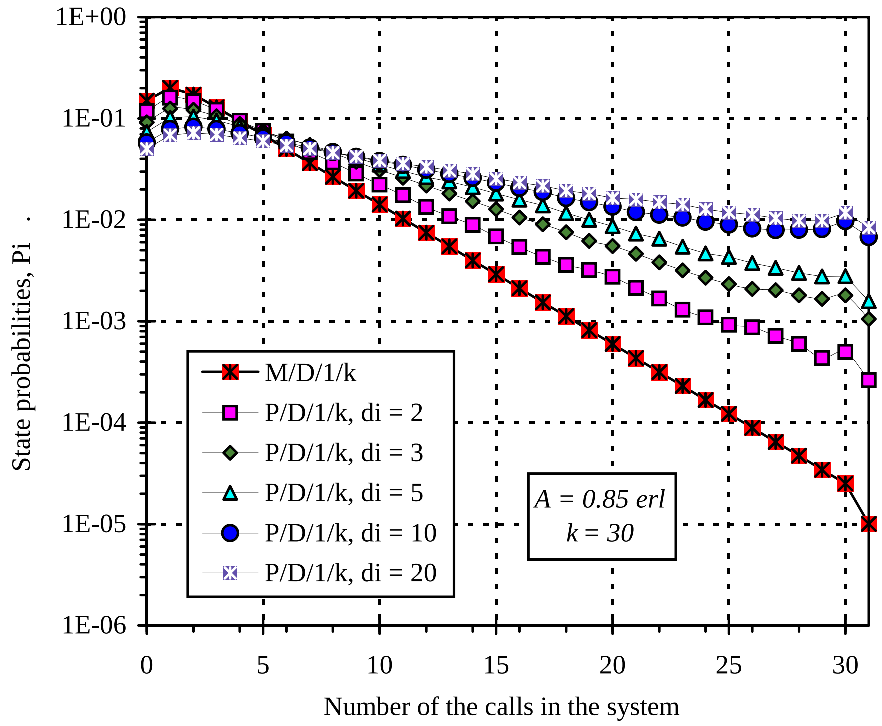

Figure 2. Stationary probability distribution of the Pareto/D/1/k when the offered traffic is A = 0.85 erl, the number of the waiting rooms *k = 30* and different peakedness of the input stream

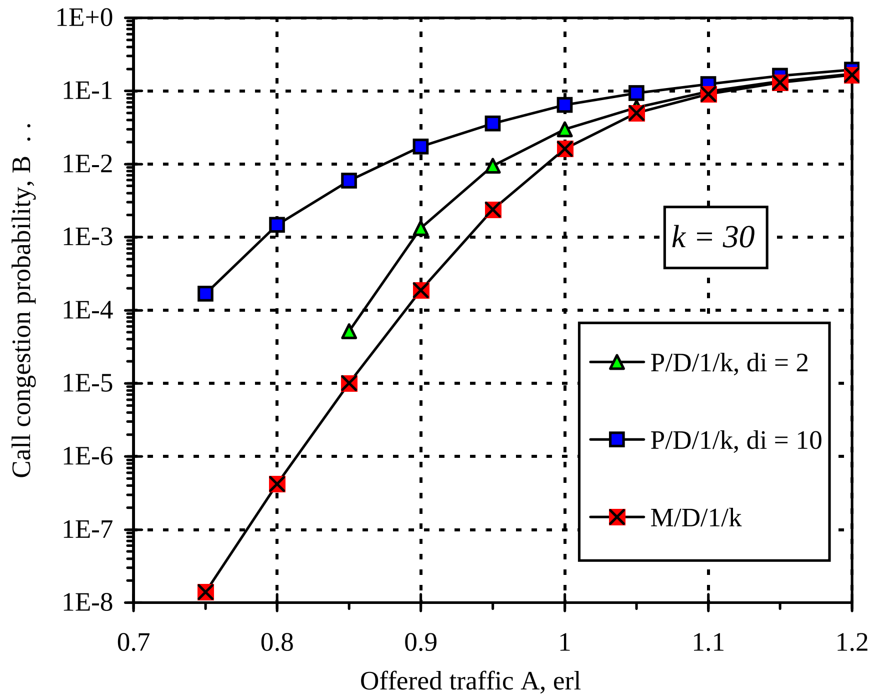

Figure 3. Call congestion probability in a single delay system with a Pareto input stream and constant service time

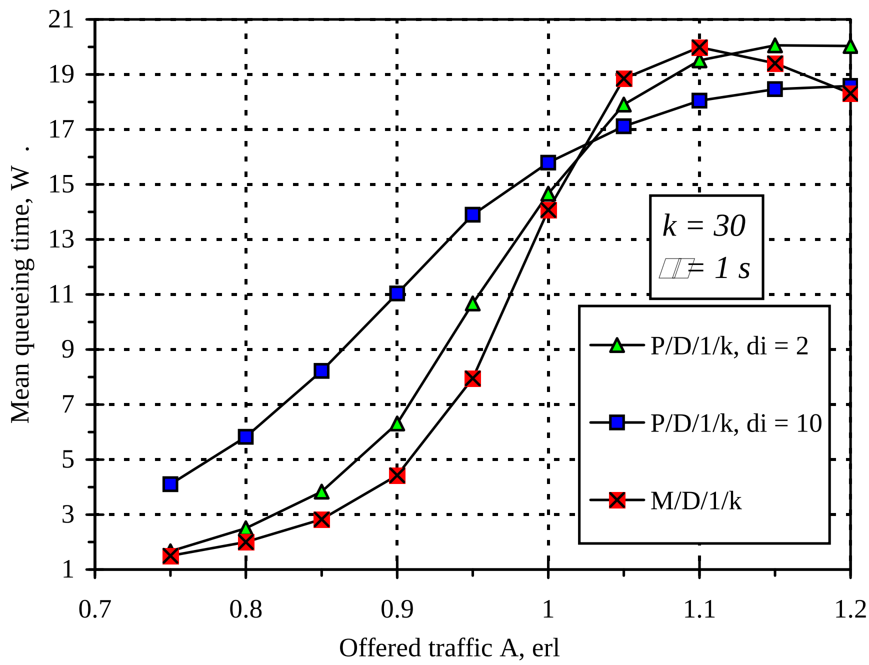

Figure 4. Mean system time in seconds in a single delay system with a Pareto input stream and constant service time

In this paper, a generalised Pareto distribution is introduced and explained. A basic simulation model for a queueing system Pareto/D/1/k is examined in detail. The developed simulation model provides a unified framework to model peak input traffic. Numerical results and subsequent experience have shown that this model is accurate and useful in analyses of teletraffic systems.

The importance of a single server queue in a case of a Pareto input stream and constant service time comes from its ability to describe behaviour that is to be found in more complex real queueing systems. It is one of the cases in a general teletraffic system that is important in telecommunication systems design.

In conclusion, we believe that the presented simulation model will be useful in practice.

Our model permits us to look at the queueing behaviour. We saw that as the load increases, the long-tailed nature of the queue brings to big losses and delay. Comparisons with Poisson arrivals showed that the simple Markovian models seriously underestimate the performance of such systems. In a sense, our results help solidify those statements being made by other authors.

The simulation method we have presented could certainly be used to study congestion in the Next Generation Networks. Our method generates a complete probabilistic analysis of the queues we study. The method is quick and its accuracy can be easy evaluated. We have used the method with the Pareto only, but are investigating its use with other distributions.

We feel that our simulation method has excellent promise to analyze the type of congestion problems and delays seen on the Internet. Thus, we are continuing our research using the simulation method for a larger class of queueing systems.

Acknowledgements

This paper is sponsored by the National Science Funds of MES - Bulgaria in the framework of project **ВУ-ТН-105/2005** “Multimedia Telecommunications Networks Planning with Quality of Service and Traffic Management”.

Bibliography

[Argibay, 2003] Argibay Losada P., A. Suárez González, C. López García, R. Rodríguez Rubio, J. López Ardao, D. Teijeiro Ruiz. On the simulation of queues with Pareto Service. Proc. 17th European Simulation Multiconference, pp. 442-447, 2003.

[Cao, 2004] Cao J., W. Cleveland, and D. Sun. Bandwidth estimation for best-effort Internet traffic. Statist. Sci., Volume 19, Number 3 (2004), pp. 518-543.

[Chlebus, 2007] Chlebus E., Gautam Divgi. The Pareto or Truncated Pareto Distribution? Measurement-Based Modelling of Session Traffic for Wi-Fi Wireless Internet Access. Wireless Communications and Networking Conference - WCNC, IEEE, pp. 3625-3630, 2007.

[Eickhoff, 2006] Eickhoff M., D. McNickle, and K. Pawlikowski. Analysis of the time evolution of quantiles in simulation. International Journal of Simulation, Vol. 7, No 6, pp. 44-55, 2006.

[Fernandes, 2003] Fernandes S., C. Kamienski, D. Sadok. Accuracy and Computational Efficiency on the Fractal Traffic Generation. Proc. of the 3rd IASTED International Multi-Conference on Wireless and Optical Communications - WOC, 2003.

[Fischer, 2005] Fischer M.; D. Bevilacqua Masi, D. Gross, J. Shorte. One-Parameter Pareto, Two-Parameter Pareto, Three-Parameter Pareto: Is There a Modelling Difference?. Telecommunications Review, pp. 79-92, 2005.

[Fischer, 1999] Fischer M.; H. Cart. A Method for Analyzing Congestion in Pareto and Related Queues. Telecommunications Review, pp. 15-27, 1999.

[Gross, 2003] Gross D., M. Fischer, D. Masi, J. Shorte. D. Gross. Difficulties in Simulating Queues with Pareto Service. Proceedings of the Winter Simulation Conference, pp. 407-415, 2003.

[Inmaculada, 2006] Inmaculada A., M. Meerschaert, and A. Panorska. Parameter Estimation for the Truncated Pareto Distribution, Journal of the American Statistical Association, Volume 101, Number 473, pp. 270-277, 2006.

[Koh, 2003] Koh Y.; Kiseon Kim. Loss probability behaviour of Pareto/M/1/K queue. Communications Letters, IEEE, Volume 7, Issue 1, pp. 39-41, Jan 2003.

[Koh, 2003] Koh Y.; Kiseon Kim. Evaluation of Steady-State Probability of Pareto/M/1/K Experiencing Tail-Raising Effect. Lecture Notes in Computer Science, Volume 2720, pp. 561-570, 2003.

[Rodrıguez, 2004] Rodrıguez-Dagnino R.. On the Pareto/M/c and Pareto/M/1/K. queues. Proc. SPIE, ITCom, vol. 5598, pp. 183-193, 2004.

[Salvador, 2004] Salvador P., A. Pacheco and R. Valadas. Modelling IP traffic: joint characterization of packet arrivals and packet sizes using BMAPs. Computer Networks, Volume 44, Issue 3 , 2004, pp. 335-352.

[Khadjiivanov, 1993] Khadjiivanov L., B.T.Taskov, A.A.Aliazidi, B.P.Tsankov. Application of Priority Queueing Mechanisms to ATM Multiplexing and Traffic Control. Proc. of Integrated Broadband Communications Networks and Services, Copenhagen, Denmark, April 20 – 23, 1993, pp. 33.3.1 – 33.3.11.

[Tsankov, 2007] Tsankov B., R. Pachamanov, and D. Pachamanova. Modified Brady Voice Traffic Model for WLAN and WMAN. Electronics Letters, vol.43, issue 23, Nov. 2007, pp. 1295-1297.

[Zhang, 2007] Zhang W.; He Jingsha. Modeling End-to-End Delay Using Pareto Distribution. Second International Conference on Internet Monitoring and Protection, ICIMP, pp. 21-21, 2007.

Authors' Information

Seferin Mirtchev – Technical University of Sofia, Kliment Ohridski St., N:8, Bl.1, Sofia-1000, Bulgaria;
e-mail: stm@tu-sofia.bg

Rossitza Goleva – Technical University of Sofia, Kliment Ohridski St., N:8, Bl.1, Sofia-1000, Bulgaria;
e-mail: rig@tu-sofia.bg